The Investigation on Birth Weight of Babies

Group 17 Hang Su, Xiaozhuo Wang, Xin Shen, Yiwei Sun

Abstract:
The purpose of this paper is to examine the relationship between maternal age and birth weight, as well as to provide the insight of how other factors affect newborn babies' birth weights. In this project, first, we select potential variables by conducting statistical tests, then build linear regression and logistic regression models based on Cp, AIC, BIC and cross validation methods, and finally select the best 2 models for birth weight prediction and low birth-weight determination. Our results show that factors such as weight gained, gestation period, plurality, marital status, maternal smoking and maturity have huge impact on causing the risk of low birth-weight. In addition, we have also demonstrated that advanced maternal age appears to contribute less risk in low birth-weights of newborns'.

1.Introduction

Birth weight is defined as body weight of a newborn at birth. In general, if a newborn baby's birth weight is smaller than 2500g, 5 pounds and 8 ounces, then it is considered as low birth weight. Low birth weight is a significant public health concern, since it is the strongest factor associated with infant mortality. In addition, the low birth weight will potentially affect health outcomes in high blood pressure, diabetes and heart disease, etc. The majority studies have shown that the risk factors associated with pregnancy are more harmful when maternal age is over 35. Therefore, we will explore the existence of any possible relationships between maternal age, external influence and birth weight. In this paper, we will have five sections, which are introduction, data set description, model I, model II and conclusion.

2.Data Set Description

The data set used in this project comes from StatCrunch website. It obtains 802 birth observations from the state of North Carolina, and consists 12 explanatory variables. The variables are provided below:

Categorical Variables
- Plurality: 0 = not twins, 1 = twins
- Baby's gender: 0 = male, 1= female
- Marital: 0 = married, 1 = single
- Smoke: 0 = no, 1 = yes
- Mature[1]: 0 = no, 1 = yes (>35 yrs)
- Race: 0 = non-black, 1= black
- Premature[2]: 0 = no, 1 = yes (born <36 wks)

Numerical Variables
- Mother's Age
- Father's Age
- Visits[3]: # of medical visits
- Gestation Period
- Weight Gained: weight gained during pregnancy

3.Model I: Linear Regression

Model Assumption:

For linear regression model, there are four assumptions:
(1) Linearity of the relationship between the response variable and the explanatory variables.
(2) Constant variance of the errors
(3) Normality of the error distribution
(4) Independence of the errors

After selecting our final model, we will conduct model checking to see whether the chosen model satisfies the assumptions of linear regression.

[1] Mature: maternal age is above 35 years or below

[2] Premature: baby is born before 36 weeks or after

[3] Visits: amount of pre-natal medical visits

Data Exploration:
Before fitting a tentative model, we need to do some data exploration to see among all of the potential explanatory variables which variables are potentially relevant to the response variable, birth weight. For each categorical variable, we first use box-plot to check whether the birth weight in each category is normally distributed or not to determine which testing method we should use to check if the mean birth weight of each category is different. The results are shown in the table below:

Variable	Method	p-value
Plurality	Rank-Sum Test	<0.01
Gender	Rank-Sum Test	<0.01
Marital	Rank-Sum Test	0.00078
Smoke	Rank-Sum Test	0.0061
Race	Rank-Sum Test	<0.01
Mature	Rank-Sum Test	0.92
Premature	Rank-Sum Test	<0.01

We can see that the p-value of mature 0.92 is significantly large, indicating the mean birth weight of babies whose mothers are over 35 is the same as the mean birth weight of babies whose mothers are below 35. We believe that mature is not relevant to the birth weight and don't include the variable in model selection step.

For the continuous variables, we check if any transformation is needed. For each continuous variable, we use three different transformations: square root, logarithmic and inverse transformation. Based on our investigation, no transformation is needed for the continuous variables and we decide to include all continuous variables in model selection step.

Model Selection:
In general, there are two reasons for model selection, to yield better prediction accuracy and to yield better model interpretability. Since we have 802 observations with only 12 potential explanatory variables, the number of observations is much larger than the number of variables. Hence, we can simply fit the full model including all the explanatory variables using the ordinary least squares fitting and still get good prediction accuracy. Saying that, our goal of model selection is for having better model interpretability. It is often the case that some or many of the variables used in a multiple regression model are in fact not associated with the response variable. Including such irrelevant variables leads to unnecessary complexity in the resulting model. Therefore, we aim to remove those variables that are not relevant to the response variable to make our final model more interpretable. The same reason is also applied to the model selection for the logistic regression part.

Then, as we know, except in special circumstances, a model including a product term for interaction between two explanatory variables should also include terms with each of the explanatory variables individually, even though their coefficients may not be significantly different from zero. Following this rule avoids the logical inconsistency of saying that the effect of X_1 depends on the level of X_2 but that there is no effect of X_1. Hence, we first try to find a tentative model without interaction terms. After finding the tentative model without interaction terms, we check whether a richer model including the interaction terms is better than the tentative model without interaction terms.

1. Best Model without interactions
Under the best subset selection approach, a number of methods including Cp statistics, BIC and 10-fold cross-validation are used to find which reduced model should be used to explain the birth weight variation.

We first use the 10-fold cross-validation and find that the best model is the one that contains 10 variables, which includes all the explanatory variables except Premature. Then, we use the Cp statistics as the criteria to choose the best model. The model with the lowest Cp statistics is the 8-variable model that contains

Plural, Gender, Mother's Age, Gestation, Marital, Weight Gained, Smoke and Race. Finally, we use the BIC as the criteria to choose the best model. The model with the lowest BIC is the 7-variable model that contains Plurality, Gender, Gestation, Marital, Weight Gained, Smoke and Race. The three selected best models are shown in the table below:

Best Reduced Model		
Method	**Model**	**SSRes**
10-Fold Cross-Validation	$-4.14-1.41\times$Plurality$-0.35\times$Gender$-0.18\times$Marital $-0.36\times$Smoke$-0.30\times$Race$+0.01\times$Father's Age $+0.01\times$Mother's Age $+0.28\times$Gestation $+0.01\times$Visits$+0.01\times$Weight Gained	861.51
Cp Statistics	$-4.10-1.39\times$Plurality$-0.35\times$Gender$-0.20\times$Marital $-0.36\times$Smoke$-0.30\times$Race$+0.01\times$Mother's Age $+0.29\times$Gestation$+0.01\times$Weight Gained	863.21
BIC	$-3.69-1.36\times$Plurality$-0.36\times$Gender$-0.27\times$Marital $-0.38\times$Smoke$-0.31\times$Race$+0.28\times$Gestation $+0.01\times$Weight Gained	867.01

We can see that the three models are actually nested, then we can use the extra-sums-of-squares F-tests to select the best model among the three models. When comparing the 10-variable model selected by the 10-fold cross-validation method and the 8-variable model selected by the Cp statistics method, the p-value 0.46 is significantly large, indicating there is no evidence that the mean birth weight is associated with father's age and the number of medical visits after accounting for other explanatory variables. When comparing the 8-variable model selected by the Cp statistics method and the 7-variable model selected by the BIC method, the p-value 0.062 is moderately small, there is suggestive but inconclusive evidence that the mean birth weight is associated with mother's age after accounting for other explanatory variables.

Therefore, we conclude, pending a model checking (found in the Model Checking section), that 8-vairable model selected by the Cp statistics method is the best model for explaining the birth weight variation.

Extra-Sums-of-Squares F-Test					
Comparison	**Extra SS**	**Number of Betas Tested**	**Estimated $\sigma^2_{\text{full model}}$**	**F-statistics**	**p-value**
Full model: 10-Fold Cross-Validation **Reduced Model:** CP Statistics	1.70	2	1.09	0.78	0.46
Full model: CP Statistics **Reduced Model:** BIC	3.80	1	1.09	3.50	0.062

2. Best Model with interactions

Now, we consider whether we should include the interaction terms between the 8 variables into the final model or not. We conduct the best subset selection using both the 10-fold cross-validation and the BIC method to determine which interaction terms, if any, should be included in the model to further explain the birth weight variation. The best models selected by the two methods are the same as shown in the table below:

Best Reduced Model		
Method	**Model**	**SSRes**
10-Fold Cross-Validation /BIC	$-7.62-1.46\times$Plurality$-0.37\times$Gender$-0.17\times$Marital $-0.37\times$Smoke$-0.29\times$Race$+0.01\times$Mother's Age $+0.38\times$Gestation$+0.15\times$Weight Gained $-0.0035\times$Gestation$\times$Weight Gained	850.19

We can see that the only interaction term needs to be included is the interaction between Gestation and Weight Gained. Then, we can use the extra-sums-of-squares F-tests to check whether the coefficient of the interaction between Gestation and Weight Gained is zero or not. Since the p-value 0.0012 is significantly small, indicating there is convincing evidence that the mean birth weight is associated with the interaction term between Gestation and Weight Gained after accounting for other explanatory variables. Therefore, we conclude, pending a model checking (found in the Model Checking section), that richer model with the interaction term between Gestation and Weight Gained is the best model for explaining the birth weight variation. We select it as our final model.

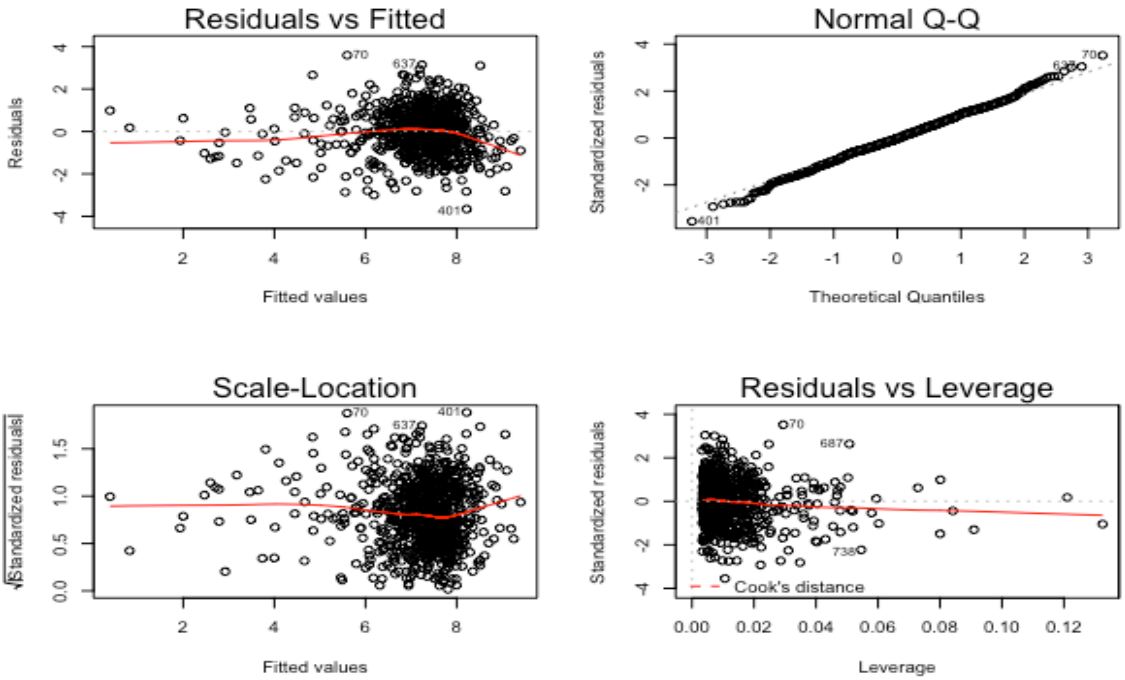
Extra-Sums-of-Squares F-Test					
Comparison	**Extra SS**	**Number of Betas Tested**	**Estimated $\sigma^2_{\text{full model}}$**	**F-statistics**	**p-value**
Full model: Only main effects **Reduced Model:** Including interactions	11.32	1	1.07	10.55	0.0012

Model Checking:
Finally, we need to check if there is any potential problem with our final model. First, we take a look at the potential multicollinearity issue. The variation inflation factor (VIF) value for each explanatory variable is shown in the table below:

VIF Value			
Plurality	**Gender**	**Marital**	**Smoke**
1.17	1.01	1.28	1.02
Race	**Mother's Age**	**Gestation**	**Weight Gained**
1.11	1.24	1.18	1.03

As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity. We can see that all the VIF values are around 1; hence, no serious collineairty issue is detected.

Next, we need to examine the residuals to check the assumptions of linear regression. Below are the plots of the final model. From the residual plot, we can see that there is no systematic pattern, suggesting that there is nonlinearity, non-constant variance or dependence of errors issue. From the Q-Q plot, we can see that there is no violation about the normality assumption.



Finally, we need to check the presence of outliers and influential observations. We first calculate the jackknife residual for each observation, and then test whether each observation is an outlier using Bonferroni correction. The largest absolute jackknife residual value is 3.57, the corresponding p-value under Bonferroni correction is 0.30 which is significant large. Hence, there is no evidence indicating the presence of outliers. Moreover, from the "Residuals vs Leverage" plot above, we can see that the Cook's distance for each observation is much smaller than 1. Hence, no influential observations are detected.

To conclude, our final model satisfies the assumption of linear regression, and no outliers or influential observations are detected. The final model is shown in the table below:

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-7.62	1.17	-6.48	<0.01
Plurality	-1.46	0.20	-7.20	<0.01
Gender	-0.37	0.074	-4.98	<0.01
Marital	-0.17	0.093	-1.85	0.065
Smoke	-0.37	0.12	-3.07	0.0022
Race	-0.29	0.098	-2.99	0.0029
Mother's Age	0.011	0.0066	1.71	0.088
Gestation	0.38	0.030	12.50	<0.01
Weight Gained	0.18	0.039	3.77	0.00017
Gestation: Weight Gained	-0.0035	0.0010	-3.48	0.00052

Form the table above, we can see that except the p-values of mother's age and marital, the p-value of each other explanatory variable is significant small (very close to 0), indicating there is convincing evidence that the mean birth weight is associated with plurality, gender, smoke, race, gestation, weight gain and the interaction between gestation and weight gain. The p-value of mother's age is 0.088 and the p-value of marital is 0.065. The two p-values are moderately small, there is suggestive but inconclusive evidence that the mean birth weight is associated with mother's age and marital. The adjusted R^2 for the final model is

0.4844 and R^2 for the final model 0.4902. Hence, 49.02% of the birth weight variation is explained by our final model.

4.Model II: Logistic Regression

Model Assumption:

The most importance thing a mother concerns is that whether her baby will have low birth weight. Thus, a logistic model is necessary to build for low weight determination. Before building a logistic model, we should check two assumptions:

(1) Linearity: we assume that the logistic model fits the data

(2) Multicollinearity: no multicollinearity .

Data Exploration:

Before fitting a tentative model, we need to do some data exploration to see among all of the potential explanatory variables which variables are potentially relevant to the binary response variable (low birth). For each categorical variable, we use the Pearson's chi-squared test to check whether the binary response variable (low birth) and the categorical variable are independent or not. The results are shown in the table below:

Pearson's Chi-squared Test P-value			
Plurality	**Gender**	**Marital**	**Smoke**
<0.01	0.98	0.0069	0.14
Race	**Mature**	**Premature**	
0.020	0.23	<0.01	

We can see that the p-value of gender 0.98, the p-value of smoke 0.14 and the p-value of mature 0.23 are significantly large; indicating the binary response variable (low birth) is independent from the categorical variable gender, smoke and mature. Hence, we don't include the three categorical variables in model selection step.

Model Selection:

Similar to the linear regression model selection, we apply the best subset selection approach based on AIC and BIC to find the best reduced model for explaining the variation of log(odds).

We first use the AIC method to find that the best model is the one that contains 4 variables, which includes Plurality, Gestation, Marital and Weight Gained. Then, we use the BIC method to choose the best model. The model with the lowest BIC is a 2-variables model that contains Plurality, Gestation. In the end, we conduct a drop-in deviance test to determine the better one. The p-value for the drop-in deviance test is smaller than 0.01, which indicates that the BIC model is not adequate and the AIC model is better. Results are shown below:

Best Reduced Model		
Method	**Model**	**Deviance**
AIC	log(odds)=20.34+2.76*Plurality-0.60*Gestation +0.91*Marital-0.033*Weight Gained	197.37
BIC	log(odds)=21.64 + 2.11 *Plurality -0.64*Gestation	207.59

Model Checking:
Finally, we need to check if there are any potential problems with our final model. We take a look at the potential multicollinearity issue. The variation inflation factor (VIF) value for each explanatory variable is shown in the table below:

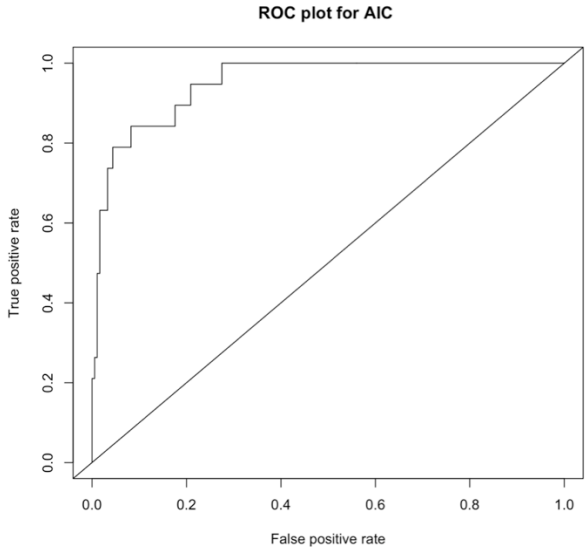
VIF Value			
Plurality	**Gender**	**Marital**	**Weight Gained**
1.148	1.018	1.059	1.089

Since all VIF values are around 1, we can assume that there are no multicollineality between variables.

To conclude, our final model satisfies the assumptions of logistic regression. The final model is shown in the table below:

AIC Model	Estimate	Std. Error	z value	Pr(>|t|)
Intercept	20.34	3.27	-6.48	5.19e-10
Plurality	2.76	0.72	3.81	1.37e-04
Gestation	-0.60	0.09	-6.75	1.46e-11
Marital	0.91	0.41	2.23	2.59e-02
Weight Gained	-0.03	0.02	-2.09	3.67e-02

Model Evaluation:
To testing how well our final model predicts results, we evaluate the performance of our logistic model on the testing data. Form the ROC plot below; we can see that our model performs much better than random guess.



Then, we need to choose a threshold for the classification. If the predicted probability of having low birth weight from our model is larger than the threshold, we classify the baby as having low birth weight. Otherwise, we classify the baby as having normal weight. Here, we set the cost of misclassification to be one for both false positive and false negative, and try to find the threshold that minimizes the total cost of misclassification. The threshold is 0.582. We classify whether a baby has low birth weight or not based on the threshold, the confusion matrix is shown below:

Prediction **True**	**1** **Low weight**	**0** **Normal weight**
1	11	3
0	8	179

Based on the confusion matrix, we can see that the true positive rate is 78.57% and the true negative rate is 95.72%. The performance is good.

5. Conclusion

In the beginning of this paper, we develop a model to predict the precise birth weight and to see what factors could contribute to the prediction. We build a linear regression model that has prediction power up to 50%, and we discover that Plurality, Baby's Gender, Mother's Age, Gestation Period, Marital, Weight Gained, Smoke, Race are relatively significant factors for birth weight prediction.

To our surprise, the marital age has a positive relationship with birth weight of babies and the coefficient is close to zero. We conjecture that this might due to the existence of teenage mothers and the lack of samples of elder mothers.

On the other hand, the logistic model works well on the testing set, giving a 5% total misclassification rate and 80% true positive rate. We believe that our model would help hospitals to predict whether the mother would give birth to a low-weight baby during pre-natal medical visits. However, the indicator of smoking turns out to have no effect on whether the baby has low birth -weight, which contradicts to our common sense. A possible explanation for this might be we miss the effect of passive smoking or other potential factors. In this sense, more detailed data will be needed to improve the model in further research.

6.Appendix

Team member contribution

Hang Su (hs2856): PPT preparation and project report
Xiaozhuo (xw2403) Wang: PPT preparation, project presentation, and project report
Xin Shen (xs2225): Linear regression and project report
Yiwei Sun (ys2882): Logistic regression and project report